JOURNAL OF COMPUTER AND SYSTEM SCIENCES 16, 301–322 (1978)

A Trivial Algorithm Whose Analysis Isn't*

ARNE T. JONASSEN† AND DONALD E. KNUTH

Computer Science Department, Stanford University, Stanford, California 94305

Received August 4, 1976; revised May 9, 1977

Very few theoretical results have been obtained to date about the behavior of information retrieval algorithms under random *deletions*, as well as random insertions. The present paper offers a possible explanation for this dearth of results, by showing that one of the simplest such algorithms already requires a surprisingly intricate analysis. Even when the data structure never contains more than three items at a time, it is shown that the performance of the standard tree search/insertion/deletion algorithm involves Bessel functions and the solution of bivariate integral equations. A step-by-step expository analysis of this problem is given, and it is shown how the difficulties arise and can be surmounted.

1. INTRODUCTION

An algorithm known as "tree search and insertion" has become one of the most commonly used methods for maintaining a dynamically growing dictionary or symbol table (see [3]). This algorithm was discovered independently by several people during the 1950s, and in 1962 Thomas N. Hibbard [1] showed that entries could also be *deleted* dynamically without difficulty. At that time Hibbard proved one of the first results that might be called a theorem of "pure computer science", because it was one of the first results ever to be proved about data structure manipulations: He showed that a random deletion from a random tree, using his algorithm, leaves a random tree. Although the statement may seem self-evident when stated in this way, it was in fact a surprising result, because the deletion algorithm was necessarily asymmetric while random trees are symmetric. Hibbard's theorem can be stated more precisely as follows: "If $n + 1$ items are inserted into an initially empty binary tree, in random order, and if one of these (selected at random) is deleted, the probability that the resulting binary tree has a given shape is the same as the probability that this tree shape would be obtained by inserting n items into an initially empty tree, in random order." It took great foresight even to

* The preparation of this paper was supported in part by the Norwegian Research Council for Science and the Humanities; by the United States National Science Foundation Grant MCS-72-03752-A03; and by the United States Office of Naval Research contract N00014-76-C-0330. Some of the calculations were performed using MACSYMA, supported by the Defense Advanced Research Projects Agency under Office of Naval Research contract N00014-75-C-0661. The U.S. Government's right to retain a nonexclusive royalty-free license in and to copyright covering this paper is acknowledged.

† Visiting Stanford University from the University of Oslo, Norway.

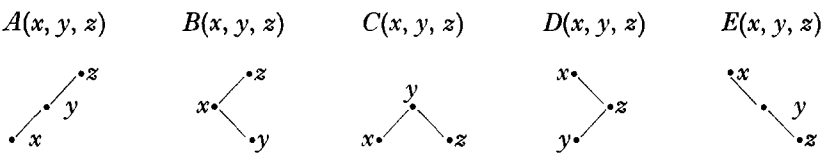
0022-0000/78/0163-0301$02.00/0

conjecture such a result in 1962; people rarely *proved* things about computer programs in those days, unless perhaps numerical analysis was involved, and binary trees were not well understood. Furthermore, the proof was not simple.

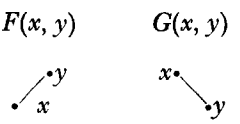
Ten years later, Gary D. Knott proved a much deeper result [2]: "If n items are inserted into an initially empty binary tree, in random order, and if the first k items inserted are subsequently deleted by Hibbard's algorithm, in the same order as they were inserted, the resulting binary tree is random." (In other words, the probability that the resulting tree has a given shape is the same as the probability that this shape of tree would be obtained if $n - k$ items had been inserted into an initially empty tree in random order.) The theorems of Hibbard and Knott seemed to settle the question of deletions, since they proved stability of the tree distribution under a wide variety of deletion disciplines.

However, Knott also discovered a surprising paradox: Although Hibbard's theorem establishes that $n + 1$ random insertions followed by a random deletion produces a tree whose shape has the distribution of n random insertions, it does *not* follow that a subsequent random insertion yields a tree whose shape has the distribution of $n + 1$ random insertions! For ten years it had been believed that Hibbard's theorem proved the stability of the algorithms under repeated insertions and deletions (cf. [1, p. 25; 3, first printing, pp. 429–432]; the discovery of a subtle fallacy in this reasoning therefore came as a shock.

In order to understand the paradox, we need to know only what Hibbard's algorithm does to binary search trees with three elements or less. The five binary search trees on three elements $x < y < z$ are



and the two possibilities on two elements $x < y$ are



The standard insertion algorithm produces the following binary search tree when inserting element z into a tree containing x and y:

Initial tree	Result if $z < x$	Result if $x < z < y$	Result if $y < z$
$F(x, y)$	$A(z, x, y)$	$B(x, z, y)$	$C(x, y, z)$
$G(x, y)$	$C(z, x, y)$	$D(x, z, y)$	$E(x, y, z)$

In other words, z is simply attached "at the bottom" where it fits. Hibbard's deletion algorithm operates as follows on a 3-element tree:

Initial tree	Delete x	Delete y	Delete z
$A(x, y, z)$	$F(y, z)$	$F(x, z)$	$F(x, y)$
$B(x, y, z)$	$F(y, z)$	$F(x, z)$	$G(x, y)$
$C(x, y, z)$	$G(y, z)$	$F(x, z)$	$F(x, y)$
$D(x, y, z)$	$G(y, z)$	$G(x, z)$	$G(x, y)$
$E(x, y, z)$	$G(y, z)$	$G(x, z)$	$G(x, y)$

If we insert three elements $x < y < z$ in random order, we get a tree of shape A, B, C, D, E with the respective probabilities 1/6, 1/6, 2/6, 1/6, 1/6; then a random deletion leaves us with the following six possibilities and probabilities:

$F(x, y)$	$F(x, z)$	$F(y, z)$	$G(x, y)$	$G(x, z)$	$G(y, z)$
3/18	4/18	2/18	3/18	2/18	4/18

The probability of shape F at this point is $9/18 = 1/2$, in accord with Hibbard's theorem.

But now comes another random insertion, say w. The probability is 1/4 that w is the smallest of $\{w, x, y, z\}$; and the other three cases $x < w < y < z$, $x < y < w < z$, $x < y < z < w$ also occur with probability 1/4. Thus the tree $F(x, y)$ becomes $A(w, x, y)$, $B(x, w, y)$ or $C(x, y, w)$ with respective probabilities 1/4, 1/4, 1/2; and the other cases $F(x, z)$,..., $G(y, z)$ can be worked out similarly. We find that the insertion of w produces a tree of shape A, B, C, D, E with the respective probabilities

$$\frac{3+4+4}{72}, \frac{3+8+2}{72}, \frac{6+4+2+3+2+8}{72}, \frac{3+4+4}{72}, \frac{6+2+4}{72},$$

namely

$$11/72, 13/72, 25/72, 11/72, 12/72. \tag{1.1}$$

A random deletion now produces a tree of shape F with probability

$$\frac{11}{72} + \frac{2}{3} \cdot \frac{13}{72} + \frac{2}{3} \cdot \frac{25}{72} = \frac{109}{216} > \frac{1}{2}.$$

A study of this example shows where the fallacy occurred: The "random" tree shape was not independent of the "random" values remaining. For example, when x is deleted (relatively large values remaining), the tree tends to be of shape G, but when z is deleted (relatively small values remaining) the tree shape is not biased towards F or G.

Fortunately the deviation from randomness occurs in the right direction here: the trees actually tend to get *better*, in the sense that the balanced shape C (which requires less

search time) becomes more probable. Extensive empirical studies by Knott [2] give overwhelming support to the conjecture that random deletions do not degrade the average search time; but no proof has yet been found.

More precisely, Knott's conjecture is this: Consider a pattern of $n + k$ insertions and n deletions, in some order, where the number of deletions never exceeds the number of insertions. For example, one of the patterns with $n = 4$ and $k = 4$ is $IIIDIIDIIIDD$.

To do each insertion, put a new random element into the tree, say a uniform random number between 0 and 1; to do each deletion, choose a random element uniformly from among those present. All of these random choices are to be independent. Then for each fixed pattern of I's and D's, the average path length of the resulting tree is conjectured to be at most equal to the average path length of the pattern consisting solely of k I's.

In attempting to explore this conjecture, it is natural to investigate the simple case of patterns

$$III, IIIDI, IIIDIDI, \ldots, III(DI)^n, \ldots$$

for $k = 3$. Such patterns never require us to deal with more than three elements in the tree at any time; so all we must do is study the following trivial procedure.

1. Let x, y be independent uniform random numbers. Insert x into an empty tree, then insert y. (If $x < y$, we get the tree $G(x, y)$, otherwise we get $F(y, x)$.)

2. Insert a new independent uniform random number into the tree.

3. Choose one of the three elements in the tree at random, each with equal probability, and delete it using Hibbard's method.

4. Return to step 2.

At the beginning of the $(n + 1)$st occurrence of step 3, we have a tree of shape A, B, C, D, or E, with certain probabilities a_n, b_n, c_n, d_n, e_n; we want to show that these probabilities approach a "steady state." According to the conjecture, c_n should be $\geqslant 1/3$, because only shape C has a path length smaller than the other shapes. The first two times we get to step 3, we have seen that $(a_n, \ldots, e_n)$ are respectively

$$\left(\frac{1}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}\right) \quad \text{and} \quad \left(\frac{11}{72}, \frac{13}{72}, \frac{25}{72}, \frac{11}{72}, \frac{12}{72}\right).$$

What do these probabilities look like after n deletions have been made, for large n? This is the problem we shall investigate in the remainder of the paper.

It turns out that this problem is not as simple as it might appear at first, in spite of the triviality of the algorithm; in fact, the analysis ranks among the more difficult of all exact analyses of algorithms that have been carried out to date, although it is "elementary" in the sense that no deep theorems of analysis are required. From the form of the answer we shall derive, it will be clear that the problem itself is *intrinsically difficult*—no really simple derivation would be able to produce such a complicated answer, and the answer is right! Since the difficulties we will encounter are interesting and instructive, an attempt

has been made to present the solution here in a motivated way, explaining how it was found, instead of simply to present a polished proof.

One might ask why the exact analysis of this process should be carried out at all, given that the answer is hard to determine; in other words, what is the point of this work? The authors first began to study the problem simply because the mathematics was challenging—surprisingly intricate yet not quite impossible—and because the problem continued to lead to interesting subproblems. The form of the final solution demonstrates that rather deep mathematics is sometimes necessary to understand programs that are very simple; and the solution procedure shows how to develop the techniques of mathematical analysis of algorithms in the new direction that was needed. However, when the problem was finally solved, the result proved to be even more interesting than expected, since the simplicity of the program combined with the difficulty of the analysis made it necessary to investigate the fundamentals of algorithmic analysis more carefully than before. The simplifications which apply in other successful analyses are missing here, so a new basic approach to studying the average behavior of algorithms using multidimensional integrals became necessary. Further work is now in progress to develop this integral-oriented approach, since it has the potential of leading to automated analysis of algorithms, extending the present techniques of automated proofs of algorithms.

2. The Recurrences to be Solved

The behavior of the trivial algorithm depends only on the relative order of the elements inserted, and on the particular choice made at each deletion step. Therefore one way to analyze the situation after the pattern $III(DI)^n$ is to consider $(n+3)!\,3^n$ configurations to be equally likely, reflecting the relative order of the $n+3$ elements inserted and the n 3-way choices of which element to delete. For example, when $n=1$ there are 72 equally likely possibilities, and our analysis of this case in (1.1) essentially considered them all.

However, such a discrete approach leads to great complications. The following continuous approach which follows the algorithm more closely turns out to be much simpler. Let $f_n(x, y)\,dx\,dy$ be the differential probability that the tree is $F(X, Y)$ at the beginning of step 2, after n elements have been deleted, where

$$x \leqslant X < x + dx \qquad \text{and} \qquad y \leqslant Y < y + dy;$$

and let $g_n(x, y)\,dx\,dy$ be the corresponding probability that it is $G(X, Y)$. Let

$$a_n(x, y, z)\,dx\,dy\,dz, \ldots, e_n(x, y, z)\,dx\,dy\,dz$$

be the respective probabilities that the tree is $A(X, Y, Z), \ldots, E(X, Y, Z)$ at the beginning of step 3, for some $x \leqslant X < x + dx$, $y \leqslant Y < y + dy$, $z \leqslant Z < z + dz$. Then it is possible to write down recurrence relations for these differential probabilities by directly translating the algorithm into mathematical formalism. First we have

$$
\begin{aligned}
a_n(x, y, z) &= f_n(y, z), \\
b_n(x, y, z) &= f_n(x, z), \\
c_n(x, y, z) &= f_n(x, y) + g_n(y, z), \\
d_n(x, y, z) &= g_n(x, z), \\
e_n(x, y, z) &= g_n(x, y), \\
&\text{for } \ 0 \leqslant x < y < z \leqslant 1,
\end{aligned} \tag{2.1}
$$

by considering the six possible actions of step 2. (These probabilities are, of course, zero when $x < 0$, $x > y$, $y > z$ or $z > 1$; at the boundaries $x = 0$, $x = y$, $y = z$, and $z = 1$ there may be discontinuities, and it does not matter how we define the functions there). Secondly we have

$$
\begin{aligned}
f_{n+1}(x, y) &= \frac{1}{3}\int_0^x (a_n(t, x, y) + b_n(t, x, y))\, dt \\
&\quad + \frac{1}{3}\int_x^y (a_n(x, t, y) + b_n(x, t, y) + c_n(x, t, y))\, dt \\
&\quad + \frac{1}{3}\int_y^1 (a_n(x, y, t) + c_n(x, y, t))\, dt, \\
g_{n+1}(x, y) &= \frac{1}{3}\int_0^x (c_n(t, x, y) + d_n(t, x, y) + e_n(t, x, y))\, dt \\
&\quad + \frac{1}{3}\int_x^y (d_n(x, t, y) + e_n(x, t, y))\, dt \\
&\quad + \frac{1}{3}\int_y^1 (b_n(x, y, t) + d_n(x, y, t) + e_n(x, y, t))\, dt, \\
&\text{for } \ 0 \leqslant x < y \leqslant 1,
\end{aligned} \tag{2.2}
$$

by considering the possible actions of step 3. Inserting (2.1) into (2.2) and applying obvious simplifications yields the fundamental recurrences

$$
\begin{aligned}
f_{n+1}(x, y) &= \frac{1}{3}\Big(f_n(x, y) + \int_0^y f_n(t, y)\, dt + \int_x^y f_n(x, t)\, dt \\
&\quad + \int_x^y g_n(t, y) + \int_y^1 f_n(y, t)\, dt + \int_y^1 g_n(y, t)\, dt\Big), \\
g_{n+1}(x, y) &= \frac{1}{3}\Big(g_n(x, y) + \int_0^x f_n(t, x)\, dt + \int_0^x g_n(t, y)\, dt \\
&\quad + \int_0^x g_n(t, x)\, dt + \int_x^1 g_n(x, t)\, dt + \int_y^1 f_n(x, t)\, dt\Big), \\
&\text{for } \ 0 \leqslant x < y \leqslant 1.
\end{aligned} \tag{2.3}
$$

Consideration of step 1 also leads to the obvious initial conditions

$$f_0(x, y) = g_0(x, y) = 1, \qquad \text{for} \quad 0 \leqslant x < y \leqslant 1. \tag{2.4}$$

We have now transformed the algorithm mechnically into a set of equations tha precisely describe the distribution of its behavior. The quantities of interest to us are

$$\begin{aligned} a_n &= \int_0^1 \int_0^z \int_0^y a_n(x, y, z)\, dx\, dy\, dz, \ldots, \\ e_n &= \int_0^1 \int_0^z \int_0^y e_n(x, y, z)\, dx\, dy\, dz, \end{aligned} \tag{2.5}$$

namely the respective probabilities that a tree of shape A,..., E occurs after the insertion/deletion pattern $III(DI)^n$; and

$$f_n = \int_0^1 \int_0^y f_n(x, y)\, dx\, dy, \qquad g_n = \int_0^1 \int_0^y g_n(x, y)\, dx\, dy, \tag{2.6}$$

the probabilities that the tree shape is F or G after the pattern $II(ID)^n$. Hibbard's theorem for trees of size 2 states that $f_0 = f_1$ and $g_0 = g_1$.

3. Simplification of the Recurrences

What can we do with such formidable recurrences (2.3)–(2.4)? In the first place we can look for invariant relations that might be used to simplify them.

When the algorithm reaches step 2, it is clear that the two numbers X and Y in its tree are random, except for the condition that $X < Y$. Thus we must have

$$f_n(x, y) + g_n(x, y) = 2, \qquad \text{for} \quad 0 \leqslant x < y \leqslant 1 \quad \text{and} \quad n \geqslant 0. \tag{3.1}$$

(It is 2, not 1, since the probability that $x \leqslant X < x + dx$ and $y \leqslant Y < y + dy$ given that $X < Y$ is $2\, dx\, dy$.) This formula could also be proved directly from (2.3) and (2.4), by induction on n.

Relation (3.1) means that we really have only one function to worry about, namely $f_n(x, y)$. Let us rewrite (2.3) and (2.4) to take account of this fact:

$$\begin{aligned} &f_0(x, y) = 1; \\ &f_{n+1}(x, y) = \frac{1}{3}\Big(2 - 2x + f_n(x, y) + \int_0^x f_n(t, y)\, dt + \int_x^y f_n(x, t)\, dt\Big), \\ &\text{for} \quad n \geqslant 0. \end{aligned} \tag{3.2}$$

Henceforth we shall avoid mentioning the condition $0 \leqslant x < y \leqslant 1$, for if we use (3.2) to define $f_n(x, y)$ for *all* x and y it will agree with the true $f_n(x, y)$ when $0 \leqslant x < y \leqslant 1$.

We have obtained a much simpler recurrence than (2.3)–(2.4), but (3.2) still has some undesirable features. Before proceeding any further, we can use (3.2) to check what we have done so far, by computing the first few f_n's:

$$f_1(x, y) = 1 - \frac{2}{3}x + \frac{1}{3}y, \qquad f_1 = \frac{1}{2};$$

$$f_2(x, y) = 1 - \frac{8}{9}x + \frac{4}{9}y + \frac{1}{18}(x - y)^2, \qquad f_2 = \frac{109}{216}.$$

Good.

We are hoping that the process converges for large n, and in this case the limiting distribution $f_\infty(x, y)$ will have to satisfy the integral equation

$$f_\infty(x, y) = \frac{1}{3}\Big(2 - 2x + f_\infty(x, y) + \int_0^x f_\infty(t, y)\, dt + \int_x^y f_\infty(x, t)\, dt\Big). \tag{3.3}$$

Before going on to find a solution to this equation, let us verify that $f_n(x, y)$ will indeed converge to $f_\infty(x, y)$ if $f_\infty(x, y)$ exists: Subtracting (3.3) from (3.2) yields

$$r_{n+1}(x, y) = \frac{1}{3}\Big(r_n(x, y) + \int_0^x r_n(t, y)\, dt + \int_x^y r_n(x, t)\, dt\Big),$$

where $r_n(x, y) = f_n(x, y) - f_\infty(x, y)$. Now if $|r_n(x, y)| \leqslant \alpha$ for $0 \leqslant x < y \leqslant 1$, we will have

$$|r_{n+1}(x, y)| \leqslant \frac{1}{3}\Big(\alpha + \int_0^x \alpha\, dt + \int_x^y \alpha\, dt\Big) = \frac{1 + y}{3}\alpha \leqslant \frac{2}{3}\alpha.$$

Therefore if $f_\infty(x, y)$ exists, so that $r_0(x, y)$ is bounded, the remainder $r_n(x, y) = O((2/3)^n)$ converges rapidly to zero, regardless of the initial distribution $f_0(x, y)$.

It remains to determine $f_\infty(x, y)$, whose defining equation (3.3) can be rewritten

$$f_\infty(x, y) = 1 - x + \frac{1}{2}\Big(\int_0^x f_\infty(t, y)\, dt + \int_x^y f_\infty(x, t)\, dt\Big). \tag{3.4}$$

The coefficient 1/2 can be removed from this relation by letting

$$q(x, y) = f_\infty(2x, 2y),$$

so that

$$q(x, y) = 1 - 2x + \int_0^x q(t, y)\, dt + \int_x^y q(x, t)\, dt. \tag{3.5}$$

What is this function $q(x, y)$? (It is suggested that the reader might enjoy trying to find it before reading on.)

4. Solving the Integral Equation

In attempting to solve (3.5), perhaps the first thing we might try is differentiation. Let $q'(x, y) = \partial q(x, y)/\partial x$, and $q_{,}(x, y) = \partial q(x, y)/\partial y$; then

$$q'(x, y) = -2 + q(x, y) + \int_x^y q'(x, t)\, dt - q(x, x), \tag{4.1}$$

$$q_{,}(x, y) = \int_0^x q_{,}(t, y)\, dt + q(x, y), \tag{4.2}$$

$$q_{,}'(x, y) = q_{,}(x, y) + q'(x, y). \tag{4.3}$$

If we postulate that q has a power series expansion

$$q(x, y) = \sum_{m,n\geqslant 0} q_{m,n} \frac{x^m}{m!} \frac{y^n}{n!}, \tag{4.4}$$

we find

$$\begin{gathered} q'(x, y) = \sum_{m,n\geqslant 0} q_{m+1,n} \frac{x^m}{m!} \frac{y^n}{n!}, \qquad q_{,}(x, y) = \sum_{m,n\geqslant 0} q_{m,n+1} \frac{x^m}{m!} \frac{y^n}{n!}, \\ q_{,}'(x, y) = \sum_{m,n\geqslant 0} q_{m+1,n+1} \frac{x^m y^n}{m!\, n!}. \end{gathered} \tag{4.5}$$

Therefore (4.3) yields the simple relation

$$q_{m+1,n+1} = q_{m,n+1} + q_{m+1,n}, \qquad \text{for} \quad m, n \geqslant 0, \tag{4.6}$$

from which it is possible to determine all the $q_{m,n}$ in terms of the boundary values $q_{0,n}$ and $q_{m,0}$.

Setting $x = 0$ in (3.5) yields

$$q(0, y) = 1 + \int_0^y q(0, t)\, dt, \tag{4.7}$$

hence $q(0, y) = e^y$ and

$$q_{0,n} = 1, \qquad \text{for} \quad n \geqslant 0. \tag{4.8}$$

Now comes a tricky manipulation, which was found while playing around trying to determine $q(x, 0)$. If we apply (4.1) with x and y interchanged, and add the two results, we get

$$\begin{aligned} q'(x, y) + q'(y, x) &= -4 + q(x, y) + q(y, x) - q(x, x) - q(y, y) \\ &\quad + \int_x^y (q'(x, t) - q'(y, t))\, dt \\ &= -4 + \int_x^y (q'(t, x) - q'(t, y))\, dt + \int_x^y (q'(x, t) - q'(y, t))\, dt. \end{aligned}$$

Let $s(x, y)$ be the symmetric function $q'(x, y) + q'(y, x)$; we have just proved that

$$s(x, y) = -4 + \int_x^y (s(x, t) - s(y, t))\, dt. \tag{4.9}$$

But this equation implies that $s(x, y) = -4$, identically! Let

$$s(x, y) = \sum_{m,n\geqslant 0} s_{m,n} \frac{x^m y^n}{m!\, n!}, \qquad s_{m,n} = q_{m+1,n} + q_{n+1,m}. \tag{4.10}$$

The coefficients $s_{m,n}$ for $m + n = k > 0$ on the left-hand side of (4.9) all arise as homogeneous linear combinations of the coefficients $s_{m,n}$ for $m + n = k - 1$, since

$$\int_x^y (x^m t^n - y^m t^n)\, dt = (x^m y^{n+1} + x^{n+1} y^m - x^{m+n+1} y^0 - x^0 y^{m+n+1})/(n + 1);$$

hence we can prove by induction on k that $s_{m,n} = 0$ whenever $m + n = k > 0$. It follows that

$$q_{m+1,n} = -q_{n+1,m}, \qquad \text{for} \quad m, n \geqslant 0 \quad \text{and} \quad m + n > 0. \tag{4.11}$$

When $m = n = 0$ we have $-4 = s_{0,0} = q_{1,0} + q_{1,0}$, hence $q_{1,0} = -2$; relations (4.6) and (4.8) imply that $q_{1,n} = n - 2$ for all $n \geqslant 0$, and (4.11) with $n = 0$ yields

$$q_{m,0} = -q_{1,m-1} = 3 - m \qquad \text{for} \quad m \geqslant 2. \tag{4.12}$$

We have found the desired boundary conditions, and it remains to deduce the general formula using (4.6). The binomial coefficient

$$\binom{m + n + a}{m + b}$$

satisfies (4.6) for all integers a and b, so it suffices to find a linear combination of these binomial coefficients, subject to the condition that the known values of $q_{m,n}$ are obtained whenever $m = 0$ or $n = 0$. The solution in this form is not unique, because of identities between binomial coefficients; probably the most elegant way to express it is

$$q_{m,n} = \binom{m + n - 3}{m} - \binom{m + n - 3}{m - 4}. \tag{4.13}$$

Our derivation has proved that $q_{m,n}$ must have this value if the power series $q(x, y)$ postulated in (4.4) satisfies (3.5). Conversely, it is clear that a power series solution to (3.5) exists, since the set of values $q_{m,n}$ with $m + n = k$ defines the set of values with $m + n = k + 1$ after integration. Therefore

$$q(x, y) = \sum_{m,n\geqslant 0} \left(\binom{m + n - 3}{m} - \binom{m + n - 3}{m - 4} \right) \frac{x^m}{m!} \frac{y^n}{n!} \tag{4.14}$$

solves (3.5). Note that $|\, q_{m,n} \,| \leqslant 2^{m+n}$, hence the power series is absolutely convergent for all x, y, and (4.14) is the only power series solution.

Finally let us try to express $q(x, y)$ in terms of simpler functions, possibly even "known" ones. The following somewhat surprising identity is especially useful for functions of this type:

$$\begin{aligned}
&e^{-x-y} \sum_{m,n\geqslant 0} \binom{m+n+a}{m+b} \frac{x^m}{m!} \frac{y^n}{n!} \\
&= \sum_{j,k,m,n\geqslant 0} (-1)^{j+k} \binom{m+n+a}{m+b} \frac{x^{j+m}}{j!\,m!} \frac{y^{k+n}}{k!\,n!} \\
&= \sum_{M,N\geqslant 0} \frac{x^M}{M!} \frac{y^N}{N!} \sum_{j,k\geqslant 0} (-1)^{j+k} \binom{M}{j}\binom{N}{k}\binom{M+N-j-k+a}{M-j+b} \\
&= \sum_{M,N\geqslant 0} \frac{x^M}{M!} \frac{y^N}{N!} \sum_{j,k\geqslant 0} (-1)^{M+b+k} \binom{M}{j}\binom{N}{k}\binom{-N+k-a+b-1}{M-j+b} \qquad (4.15) \\
&= \sum_{M,N\geqslant 0} \frac{x^M}{M!} \frac{y^N}{N!} \sum_{k\geqslant 0} (-1)^{M+b+k} \binom{N}{k}\binom{M-N+k-a+b-1}{M+b} \\
&= \sum_{M,N\geqslant 0} \frac{x^M}{M!} \frac{y^N}{N!} (-1)^{M+N+b} \binom{M-N-a+b-1}{M-N+b} \\
&= \sum_{M,N\geqslant 0} \frac{x^M}{M!} \frac{y^N}{N!} \binom{a}{M-N+b}.
\end{aligned}$$

When $M - N$ has a fixed value, the terms of this sum are readily expressed in terms of modified Bessel functions of the first kind, defined as usual by the formula

$$I_r(2z) = \sum_{\substack{k\geqslant 0 \\ k\geqslant -r}} \frac{z^{2k+r}}{k!\,(k+r)!}. \qquad (4.16)$$

For example, if $a \geqslant 0$ all terms vanish except those for $0 \leqslant M - N + b \leqslant a$, hence (4.15) reduces to a finite sum

$$\sum_r \binom{a}{r} \sum_{\substack{M,N\geqslant 0 \\ M+b=N+r}} \frac{x^M}{M!} \frac{y^N}{N!} = \sum_r \binom{a}{r} \left(\left(\frac{x}{y}\right)^{1/2}\right)^{r-b} I_{r-b}(2(xy)^{1/2}).$$

On the other hand, if g < 0 (as it unfortunately is in our case), another function is apparently required.

Let $h(x, y)$ be the double power series

$$\sum_{m\geqslant n\geqslant 0} \frac{x^m}{m!} \frac{y^n}{n!} \qquad (4.17)$$

which converges absolutely for all x and y. We have

$$h(x, y) = \sum_{m,n\geqslant 0} \frac{x^{m+n}}{m!}\frac{y^n}{n!} = \sum_{m\geqslant 0} \left(\left(\frac{x}{y}\right)^{1/2}\right)^m I_m(2(xy)^{1/2}). \tag{4.18}$$

Furthermore

$$\begin{aligned} h(x, y) &= e^y \sum_{m\geqslant 0} \frac{x^m}{m!}\left(1 - \frac{1}{m!}\int_0^y e^{-t}t^m\,dt\right) \\ &= e^{x+y} - e^y \int_0^y e^{-t}\left(\sum_{m\geqslant 0} \frac{t^m x^m}{m!\,m!}\right) dt \\ &= e^{x+y} - e^y \int_0^y e^{-t} I_0(2(tx)^{1/2})\,dt, \end{aligned} \tag{4.19}$$

so $h(x, y)$ can be expressed in at least two ways in terms of Bessel functions; but it does not seem to have any simpler expressions in "closed form". The definition of $h(x, y)$ is already sufficiently simple that we can consider it a known function; we will express $q(x, y)$ in terms of $h(x, y)$ and Bessel functions.

By (4.14) and (4.15),

$$\begin{aligned} e^{-x-y}q(x, y) &= \sum_{m,n\geqslant 0} \frac{x^m}{m!}\frac{y^n}{n!}(-1)^{m+n}\left(\binom{m-n+2}{m-n} - \binom{m-n-2}{m-n-4}\right) \\ &= \sum_{m\geqslant n\geqslant 0} \frac{x^m}{m!}\frac{y^n}{n!}(-1)^{m+n}(4m - 4n - 2 + 3\delta_{m,n} + \delta_{m,n+1}) \\ &= 4xyi_1(xy) - 4xh(-x, -y) + 4yh(-x, -y) - 4yi_0(xy) \\ &\quad - 2h(-x, -y) + 3i_0(xy) - xi_1(xy) \end{aligned}$$

where $i_r(z) = \sum_{k\geqslant 0} z^k/k!(k + r)!$. This yields the steady-state distribution $f_\infty(x, y)$ of the trivial algorithm, if we replace x and y by $x/2$ and $y/2$:

$$\begin{aligned} f_\infty(x, y) &= e^{(x+y)/2}\left((2y - 2x - 2)\,h\left(-\frac{x}{2}, -\frac{y}{2}\right) + (3 - 2y)\,I_0((xy)^{1/2})\right. \\ &\quad \left. + \frac{(2y - 1)\,x}{(xy)^{1/2}} I_1((xy)^{1/2})\right), \\ &\text{for } \quad 0 \leqslant x < y \leqslant 1. \end{aligned} \tag{4.20}$$

5. An Explicit Formula for $f_n(x, y)$

Now that the limiting behavior has been found, we can look back at the original recurrence (3.2) and see that it does not appear so formidable any more. Let us define a sequence of polynomials as follows:

$$p_0(x, y) = 1, \tag{5.1}$$

$$p_1(x, y) = y - 2x, \tag{5.2}$$

$$p_{k+1}(x, y) = \int_0^x p_k(t, y)\, dt + \int_x^y p_k(x, t)\, dt, \qquad \text{for} \quad k \geqslant 1. \tag{5.3}$$

Thus $p_2(x, y) = (1/2)(x - y)^2$, $p_3(x, y) = (1/6)y^3$, etc.; it is easy to see that each term of $p_k(x, y)$ has total degree k.

These polynomials handle the complicated parts of recurrence (3.2). If we assume that $f_n(x, y)$ is a linear combination of the p's, say

$$f_n(x, y) = \sum_{k \geqslant 0} \varphi_{n,k} \cdot p_k(x, y) \tag{5.4}$$

with $\varphi_{n,0} = 1$, relations (3.2) and (5.3) imply that $f_{n+1}(x, y)$ also has such a representation, namely

$$\begin{aligned} f_{n+1}(x, y) &= \frac{1}{3}\Big(2 - 2x + f_n(x, y) + y + \sum_{k \geqslant 1} \varphi_{n,k} p_{k+1}(x, y)\Big) \\ &= 1 + \frac{1}{3}\Big(\sum_{k \geqslant 1} \varphi_{n,k} p_k(x, y) + \sum_{k \geqslant 0} \varphi_{n,k} p_{k+1}(x, y)\Big). \end{aligned}$$

Hence (5.4) holds for all n if the coefficients $\varphi_{n,k}$ satisfy

$$\begin{aligned} \varphi_{n+1,0} &= 1, \\ \varphi_{n+1,k+1} &= (1/3)(\varphi_{n,k+1} + \varphi_{n,k}), \qquad \text{for} \quad n \geqslant 0 \quad \text{and} \quad k \geqslant 0. \end{aligned} \tag{5.5}$$

Since $\varphi_{0,k} = 0$ for all $k \geqslant 1$, this recurrence is easy to solve, and we have

$$\varphi_{n,k} = \sum_{1 \leqslant j \leqslant n} \binom{j-1}{k-1} 3^{-j}, \qquad \text{for} \quad n \geqslant 0 \quad \text{and} \quad k \geqslant 1. \tag{5.6}$$

Equation (5.4) would now be a fairly explicit formula for $f_n(x, y)$, if we only knew $p_k(x, y)$.

Let $n \to \infty$; then

$$\varphi_{\infty,k} = \sum_{j \geqslant 1} \binom{j-1}{k-1} 3^{-j} = 2^{-k}, \qquad \text{for} \quad k \geqslant 1. \tag{5.7}$$

Since $f_\infty(2x, 2y) = q(x, y)$, and since all terms of $p_k(x, y)$ have total degree k, we must have

$$q(x, y) = \sum_{k \geqslant 0} p_k(x, y). \tag{5.8}$$

Therefore we can find $p_k(x, y)$ by selecting the terms of total degree k in (4.14), namely

$$p_k(x, y) = \frac{1}{k!} \sum_j \binom{k}{j}\left(\binom{k-3}{j} - \binom{k-3}{j-4}\right) x^j y^{k-j}. \tag{5.9}$$

We may also express $p_k(x, y)$ in "closed form", in terms of the Jacobi polynomials defined by

$$(x-y)^n P_n^{(\alpha,\beta)}\left(\frac{x+y}{x-y}\right) = \sum_j \binom{n+\alpha}{j}\binom{n+\beta}{n-j} x^j y^{n-j}; \tag{5.10}$$

the result is

$$p_k(x, y) = \frac{1}{k!}\left((x-y)^k P_k^{(-3,0)}\left(\frac{x+y}{x-y}\right) - x^4(x-y)^{k-4} P_{k-4}^{(1,4)}\left(\frac{x+y}{x-y}\right)\right). \tag{5.11}$$

6. Approach to the Answers

We have shown that the trivial algorithm leads to a (nontrivial) limiting distribution. What we really want to know is the limiting probabilities of the various tree shapes that arise, namely the quantities $a_n, \ldots, e_n, f_n$, and g_n defined by the integrals in (2.5) and (2.6), as $n \to \infty$.

We clearly have

$$a_n + b_n + c_n + d_n + e_n = 1, \tag{6.1}$$

$$f_n + g_n = 1. \tag{6.2}$$

Furthermore since $b_n(x, y, z) + d_n(x, y, z) = 2$ by (2.1) and (3.1), we have

$$b_n + d_n = 1/3. \tag{6.3}$$

Another relation, slightly more subtle, also holds. We have

$$a_n = \iiint_{0\leqslant x\leqslant y\leqslant z\leqslant 1} f_n(y, z)\, dx\, dy\, dz = \iint_{0\leqslant x\leqslant y\leqslant 1} x f_n(x, y)\, dx\, dy,$$

$$b_n = \iiint_{0\leqslant x\leqslant y\leqslant z\leqslant 1} f_n(x, z)\, dx\, dy\, dz = \iint_{0\leqslant x\leqslant y\leqslant 1} (y-x) f_n(x, y)\, dx\, dy,$$

$$\frac{1}{3} - e_n = \iiint_{0\leqslant x\leqslant y\leqslant z\leqslant 1} f_n(x, y)\, dx\, dy\, dz = \iint_{0\leqslant x\leqslant y\leqslant 1} (1-y) f_n(x, y)\, dx\, dy.$$

Therefore

$$a_n + b_n + 1/3 - e_n = f_n. \tag{6.4}$$

And still another relation, even more subtle, can be obtained by looking more closely. If we integrate both sides of (3.2) over $0 \leqslant x \leqslant y \leqslant 1$ we find

$$\begin{aligned} 3f_{n+1} &= \frac{2}{3} + f_n + \iint_{0\leqslant x\leqslant y\leqslant 1} \int_0^x f_n(t, y)\, dt + \iint_{0\leqslant x\leqslant y\leqslant 1} \int_x^y f_n(x, t)\, dt \\ &= \frac{2}{3} + f_n + b_n + \frac{1}{3} - e_n. \end{aligned}$$

Combining this with (6.4) yields the somewhat surprising formula

$$a_n + 3f_{n+1} = 2/3 + 2f_n\,. \tag{6.5}$$

For example, we know that $a_1 = 11/72$, $f_1 = 1/2$, and $f_2 = 109/216$; everything checks out beautifully.

From relations (6.1)–(6.5) we can determine all of $a_n, \ldots, e_n, f_n, g_n$ knowing only the values of b_n and f_n for all n. Let us first look at f_n, and especially at the component involving $p_k(x, y)$:

$$\begin{aligned}\iint_{0\leqslant x\leqslant y\leqslant 1} p_k(x,y)\,dx\,dy &= \frac{1}{k!}\sum_j \binom{k}{j}\left(\binom{k-3}{j} - \binom{k-3}{j-4}\right)\frac{1}{j+1}\,\frac{1}{k+2}\\ &= \frac{1}{(k+2)!}\sum_j \binom{k+1}{j+1}\left(\binom{k-3}{j} - \binom{k-3}{j-4}\right)\\ &= \frac{1}{(k+2)!}\left(\binom{2k-2}{k} - \binom{2k-2}{k-4}\right).\end{aligned} \tag{6.6}$$

Similarly

$$\begin{aligned}\iint_{0\leqslant x\leqslant y\leqslant 1} (y-x)\,p_k(x,y)\,dx\,dy &= \frac{1}{k!}\sum_j \binom{k}{j}\left(\binom{k-3}{j} - \binom{k-3}{j-4}\right)\frac{1}{j+1}\,\frac{1}{j+2}\,\frac{1}{k+3}\\ &= \frac{1}{(k+3)!}\sum_j \binom{k+2}{j+2}\left(\binom{k-3}{j} - \binom{k-3}{j-4}\right)\\ &= \frac{1}{(k+3)!}\left(\binom{2k-1}{k} - \binom{2k-1}{k-4}\right).\end{aligned} \tag{6.7}$$

These quantities are nonnegative for all $k \geqslant 0$, and since the coefficients $\varphi_{n,k}$ in (5.4) and (5.6) are monotone nondecreasing with n, it follows that

$$f_{n+1} \geqslant f_n \qquad \text{and} \qquad b_{n+1} \geqslant b_n \qquad \text{for} \quad n \geqslant 0. \tag{6.8}$$

(A similar argument shows that $e_{n+1} \leqslant e_n$ for all n.)

Let us now look at the limiting behavior. We have

$$f_\infty(x,y) = \sum_{k\geqslant 0} \frac{1}{2^k}\,p_k(x,y)$$

by (5.7), hence by (6.6) and (6.7) the probabilities f_n and b_n increase to the limits

$$\begin{aligned} f_\infty &= \sum_{k\geqslant 0} \frac{1}{2^k(k+2)!}\left(\binom{2k-2}{k} - \binom{2k-2}{k-4}\right),\\ b_\infty &= \sum_{k\geqslant 0} \frac{1}{2^k(k+3)!}\left(\binom{2k-1}{k} - \binom{2k-1}{k-4}\right).\end{aligned} \tag{6.9}$$

7. Evaluation of the Final Sums

The formulas in (6.9) converge rapidly, so we could compute them and be done; but of course we would like to express the result in terms of "known" mathematical quantities, for if there is a simple answer we want to know about it. In order to get a cleaner sum to work with, let us consider the similar series

$$s_r(x) = \sum_{k\geqslant 0} \frac{(x/2)^k}{(k+r)!}\binom{2k}{k} \tag{7.1}$$

which converges absolutely for all x. Differentiation yields

$$\begin{aligned} s_r'(x) &= \sum_{k\geqslant 1} \frac{(x/2)^{k-1}}{(k+r)!}(2k-1)\binom{2k-2}{k-1} \\ &= \sum_{k\geqslant 0} \frac{(x/2)^k}{(k+r+1)!}(2k+1)\binom{2k}{k} \\ &= \sum_{k\geqslant 0} \frac{(x/2)^k}{(k+r+1)!}(2(k+r+1)-(2r+1))\binom{2k}{k} \\ &= 2s_r(x) - (2r+1)\,s_{r+1}(x). \end{aligned} \tag{7.2}$$

Thus if we define

$$t_r(x) = e^{-2x}s_r(x), \tag{7.3}$$

we have

$$t_r'(x) = -(2r+1)\,t_{r+1}(x). \tag{7.4}$$

According to this relation, we obtain all $t_r(x)$ by starting with $t_0(x)$ and differentiating.

A curious thing happens when we look at $t_0(x)$:

$$\begin{aligned} e^{-2x}s_0(x) &= \sum_{k\geqslant 0} \frac{(x/2)^k}{k!}\binom{2k}{k}\sum_{j\geqslant 0}\frac{(-2x)^j}{j!} \\ &= \sum_{m\geqslant 0}\frac{(-2x)^m}{m!}\sum_k \binom{m}{k}\frac{(-1)^k}{4^k}\binom{2k}{k} \\ &= \sum_{m\geqslant 0}\frac{(-2x)^m}{m!}\sum_k \binom{m}{m-k}\binom{-1/2}{k} \\ &= \sum_{m\geqslant 0}\frac{(-2x)^m}{m!}\binom{m-1/2}{m} \\ &= \sum_{m\geqslant 0}\frac{(-x/2)^m}{m!}\binom{2m}{m} = s_0(-x), \end{aligned}$$

using the familiar identities

$$(-1)^m \binom{-1/2}{m} = \binom{m-1/2}{m} = 4^{-m}\binom{2m}{m}. \tag{7.5}$$

In other words, $t_0(x) = s_0(-x)$, and $e^{-x}s_0(x) = e^x s_0(-x)$ is an *even* function! This coincidence deserves looking into; let us write

$$\begin{aligned} e^{-x}s_0(x) &= \sum_{k\geqslant 0} \frac{(x/2)^k}{k!}\binom{2k}{k} \sum_{j\geqslant 0}\frac{(-x)^j}{j!} \\ &= \sum_{m\geqslant 0}\frac{(-x)^m}{m!}u_m \end{aligned}$$

where

$$u_m = \sum_k \binom{m}{k}\frac{(-1)^k}{2^k}\binom{2k}{k}. \tag{7.6}$$

After a few moments of playing with this sum, an experienced binomial-coefficientologist might hit on the following elementary method of evaluation:

$$\begin{aligned} u_m &= u_{m-1} + \sum_k \binom{m-1}{k-1}\frac{(-1)^k}{2^k}\binom{2k}{k} \\ &= u_{m-1} - \sum_k \binom{m-1}{k}\frac{(-1)^k}{2^{k+1}}\binom{2k+2}{k+1} \\ &= u_{m-1} - \sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k}\frac{2k+1}{m} \\ &= u_{m-1} - \sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k}\frac{2k+2}{m} + \sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k}\frac{1}{m} \\ &= u_{m-1} - 2u_{m-1} + \frac{1}{m}\sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k}, \end{aligned}$$

hence

$$m(u_m + u_{m-1}) = \sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k},$$

$$(m-1)(u_{m-1} + u_{m-2}) = \sum_k \binom{m-1}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k} = \sum_k \binom{m}{k+1}\frac{(-1)^k}{2^k}\binom{2k}{k} - u_{m-1}.$$

Subtracting these equations yields

$$mu_m = (m-1)u_{m-2}.$$

Now $u_0 = 1$ and $u_1 = 0$, hence $u_{2m+1} = 0$, as we knew; and

$$u_{2m} = \frac{2m-1}{2m}\frac{2m-3}{2m-2}\cdots\frac{1}{2} = \binom{m-1/2}{m} = 4^{-m}\binom{2m}{m}. \tag{7.7}$$

(Is there a simpler elementary proof of this formula?) We have shown that

$$e^{-x}s_0(x) = \sum_{m\geqslant 0}\frac{(-x)^{2m}}{(2m)!}u_{2m} = \sum_{m\geqslant 0}\frac{(x/2)^{2m}}{m!\,m!} = I_0(x); \tag{7.8}$$

so our friend the modified Bessel function has appeared again. The above relations now yield the identities

$$s_r(x) = e^{2x}t_r(x) = e^{2x}\frac{(-1)^r}{1\cdot 3\cdot\ldots\cdot(2r-1)}\frac{d^r}{dx^r}(e^{-x}I_0(x)),$$

so that

$$\begin{aligned} s_0(x) &= e^x I_0(x),\\ s_1(x) &= e^x(I_0(x) - I_0'(x)),\\ s_2(x) &= \frac{1}{3}e^x(I_0(x) - 2I_0'(x) + I_0''(x)),\\ s_3(x) &= \frac{1}{15}e^x(I_0(x) - 3I_0'(x) + 3I_0''(x) - I_0'''(x)), \quad \text{etc.} \end{aligned} \tag{7.9}$$

It is easy to see from definition (4.16) that

$$I_0'(x) = I_1(x), \qquad I_1'(x) = I_0(x) - x^{-1}I_1(x), \tag{7.10}$$

hence we can express each $s_r(x)$ in terms of $I_0(x)$ and $I_1(x)$.

Finally to get f_∞ and b_∞ we need to express the sums in (6.9) in terms of $s_r(x)$ for various r. The problem boils down to expressing the binomial coefficient $\binom{2n+m}{n}$ as a linear combination of binomial coefficients of the form $\binom{2n+2k}{n+k}$. For $m = 0$ this is no problem, and for $m = 1$ we have

$$\binom{2n+1}{n} = \frac{1}{2}\binom{2n+2}{n+1} \qquad \text{if} \quad n \geqslant 0.$$

For $m \geqslant 2$ we can reduce the prblem to the cases $m - 1$ and $m - 2$, since

$$\binom{2n+m}{n} = \binom{2n+2+(m-1)}{n+1} - \binom{2n+2+(m-2)}{n+1}.$$

Iterating this idea leads us to the desired identity,

$$\binom{2n+m}{n} = \frac{1}{2}\sum_{0<k\leqslant m}(-1)^{m-k}\binom{2n+2k}{n+k}\binom{k}{m-k}\frac{m}{k}, \qquad \text{for} \quad m \geqslant 1, \quad n \geqslant -m/2. \tag{7.11}$$

In particular we get

$$\binom{2n-2}{n} = \binom{2n-2}{n-2} = \frac{1}{2}\binom{2n}{n} - \binom{2n-2}{n-1} + \frac{1}{2}\delta_{n,0},$$

$$\binom{2n-2}{n-4} = \frac{1}{2}\binom{2n+4}{n+2} - 3\binom{2n+2}{n+1} + \frac{9}{2}\binom{2n}{n} - \binom{2n-2}{n-1} - \frac{3}{2}\delta_{n,0},$$

$$\binom{2n-1}{n} = \frac{1}{2}\binom{2n}{n} + \frac{1}{2}\delta_{n,0},$$

$$\binom{2n-1}{n-4} = \frac{1}{2}\binom{2n+6}{n+3} - \frac{7}{2}\binom{2n+4}{n+2} + 7\binom{2n+2}{n+1} - \frac{7}{2}\binom{2n}{n} + \frac{1}{2}\delta_{n,0}$$

for $n \geqslant 0$.

Letting s_r stand for $s_r(1)$, we can now rewrite (6.9) as

$$\begin{aligned} f_\infty &= \frac{1}{2}s_2 - \frac{1}{2}s_3 - \frac{4}{2}(s_0 - 1 - 1) + 6(s_1 - 1) - \frac{9}{2}s_2 + \frac{1}{2}s_3 + 1 \\ &= -1 - 2s_0 + 6s_1 - 4s_2 \\ &= \frac{4}{3}eI_0(1) - 2eI_1(1) - 1; \\ b_\infty &= \frac{1}{2}s_3 - \frac{8}{2}\left(s_0 - 1 - 1 - \frac{3}{4}\right) + \frac{7\cdot 4}{2}\left(s_1 - 1 - \frac{1}{2}\right) - 7\cdot 2\left(s_2 - \frac{1}{2}\right) + \frac{7}{2}s_3 \\ &= -3 - 4s_0 + 14s_1 - 14s_2 + 4s_3 \\ &= 2eI_0(1) - \frac{12}{5}eI_1(1) - 3. \end{aligned} \tag{7.12}$$

The Bessel function values we need are readily computed to be

$$\begin{aligned} I_0(1) &= 1.26606\ 58777\ 52008\ 33559\ 82446\ 25214\ 71753\ 76077 -, \\ I_1(1) &= 0.56515\ 91039\ 92485\ 02720\ 76960\ 27609\ 86330\ 73289 -. \end{aligned} \tag{7.13}$$

Finally therefore we have the answers:

$$\begin{aligned} a_\infty &= 2/3 - f_\infty &&= 0.15049\ 16196\ 41488\ 77320, \\ b_\infty & &&= 0.19601\ 96040\ 80347\ 57536, \\ c_\infty &= f_\infty - e_\infty &&= 0.35250\ 55369\ 95186\ 10505, \\ d_\infty &= 1/3 - b_\infty &&= 0.13731\ 37292\ 52985\ 75797, \\ e_\infty &= 1 + b_\infty - 2f_\infty &&= 0.16366\ 95100\ 29991\ 78842, \\ f_\infty & &&= 0.51617\ 50470\ 25177\ 89347, \\ g_\infty &= 1 - f_\infty &&= 0.48382\ 49529\ 74822\ 10653. \end{aligned} \tag{7.14}$$

The average internal path length of the tree just before the $(n+1)$st deletion is $3a_n + 3b_n + 2c_n + 3d_n + 3e_n = 3 - c_n$. We have proved that c_n converges to c_∞, which is greater than $c_0 = 1/3$; this is consistent with the conjecture that deletions do not make the path length larger than pure insertions do. However, it is interesting to note that the convergence of c_n to c_∞ is *not* monotonic:

$$c_0 = \frac{1}{3} = 0.33333$$

$$c_1 = \frac{25}{72} = 0.34722$$

$$c_2 = \frac{19}{54} = 0.35185$$

$$c_3 = \frac{143}{405} = 0.35309$$

$$c_4 = \frac{3004}{8505} = 0.35320$$

$$c_5 = \frac{1152983}{3265920} = 0.35303$$

$$c_6 = \frac{4667107}{13226976} = 0.35285$$

$$c_7 = \frac{699791131}{1984046400} = 0.35271.$$

Therefore random deletions do *not* always enhance the average path length; the pattern *IIIDIDIDIDI* leads to a better average search time than does the same pattern followed by *DI*, and an argument that does not rely on such monotonicity will be necessary to prove Knott's conjecture.

8. Modified Deletions

To complete our study of this process we should also look at what happens if the "improved" deletion algorithm discussed in [3, p. 432] is used. Here a new "step $D1\frac{1}{2}$" is introduced, to simplify the deletion of nodes having an empty left subtree.

The modified algorithm changes only one thing with respect to trees with three or fewer nodes: the deletion of x from $D(x, y, z)$ now produces $F(y, z)$ instead of $G(y, z)$. The net effect is that the integral

$$\int_0^x g_n(t, y)\, dt$$

moves from the sum for $g_{n+1}(x, y)$ to the sum for $f_{n+1}(x, y)$ in (2.3).

Fortunately this change makes the analog of (3.2) much simpler than before; we now have

$$\begin{aligned} f_0(x, y) &= 1 \\ f_{n+1}(x, y) &= \frac{1}{3}\Big(2 + f_n(x, y) + \int_x^y f_n(x, t)\, dt\Big) \\ &\text{for} \quad n \geqslant 0, \end{aligned} \tag{8.1}$$

since (3.1) remains valid. The relation corresponding to (3.3) reduces to

$$f_\infty(x, y) = 1 + \frac{1}{2}\int_x^y f_\infty(x, t)\, dt, \tag{8.2}$$

and by arguing as before (but with considerably fewer complications) we can deduce the solution

$$f_\infty(x, y) = e^{(y-x)/2}. \tag{8.3}$$

In fact, it is not difficult to establish the general formula

$$f_n(x, y) = \sum_{0 \leqslant k \leqslant n} \frac{(y-x)^k}{k!} \sum_{k \leqslant t \leqslant n} \left(\frac{1}{3}\right)^t \binom{t-1}{t-k} \qquad \text{for} \quad n \geqslant 0. \tag{8.4}$$

Since $f_\infty(x, y)$ now has such a simple form, we can easily determine the limiting integrals corresponding to (2.5) and (2.6):

$$\begin{aligned} a_\infty &= 8e^{1/2} - 13 &&= 0.1897701\ldots, \\ b_\infty &= 20 - 12e^{1/2} &&= 0.2153447\ldots, \\ c_\infty &= 1/3 &&= 0.3333333\ldots, \\ d_\infty &= 1/3 - b_\infty &&= 0.1179885\ldots, \\ e_\infty &= 1/3 - a_\infty &&= 0.1435631\ldots, \\ f_\infty &= 4e^{1/2} - 6 &&= 0.5948850\ldots, \\ g_\infty &= 7 - 4e^{1/2} &&= 0.4051149\ldots. \end{aligned} \tag{8.5}$$

As expected, there is now a stronger bias towards the F tree. The unexpected result is that c_∞ has such a simple form compared to the others; in fact it turns out that

$$c_n = 1/3 \qquad \text{for all} \quad n \geqslant 0, \tag{8.6}$$

so the average internal path length is the same as that of a random tree built up from three insertions! Equation (8.6) follows easily from (8.4) and the fact that

$$\iiint_{0 \leqslant x \leqslant y \leqslant z \leqslant 1} ((y-x)^k - (z-y)^k)\, dx\, dy\, dz = 0 \qquad \text{for} \quad k \geqslant 0.$$

Since the values of c_n in the unmodified algorithm are *greater* than 1/3, for $n \geqslant 1$, the average internal path length actually turns out to be worse when we use the "improved" algorithm. On the other hand, Knott's empirical data in [2] indicate that the modified algorithm does indeed lead to an improvement when the trees are larger.

References

1. T. N. Hibbard, Some combinatorial properties of certain trees with applications to searching and sorting, *J. Assoc. Comput. Mach.* **9** (1962), 13–28.
2. G. D. Knott, "Deletion in Binary Storage Trees," Ph. D. Thesis, Computer Science Department, Stanford University, May 1975.
3. D. E. Knuth, "The Art of Computer Programming," Vol. 3, "Sorting and Searching,", 2nd printing, rev. Sect. 6.2.2., Addison–Wesley, Reading, Mass., March 1975.
4. D. E. Knuth, Deletions that preserve randomness, *IEEE Trans. Software Eng.* **SE-3** (1977), 351–359.